"WHY AMD FOR HPC" [CPU]

JASON HOGAN-O'NEILL, DIRECTOR HPC CENTRE OF EXCELLENCE

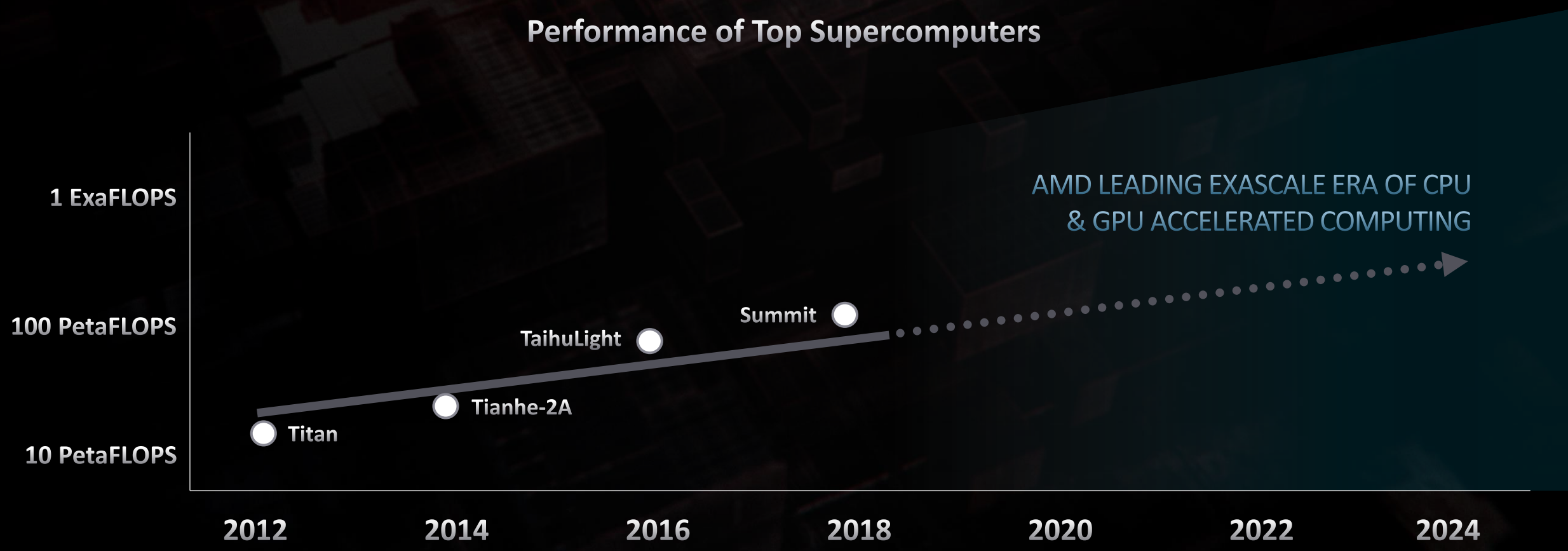
ADVANCED MODELLING AND SIMULATION SEMINAR SERIES
NASA AMES RESEARCH CENTER , JULY 20TH 2021

OVERVIEW

- Introduction. Pitch on “Why AMD for CPU”
- SKU orientation
- Architecture
- Software Development Environment
  - SPACK HPC Package Management
- Applications and their Characterisation
- References

ACCELERATED COMPUTING

TACKLING THE WORLD'S MOST IMPORTANT COMPUTATIONAL CHALLENGES



Source: https://openai.com/blog/ai-and-compute/ (Machine Intelligence) and https://www.top500.org/ (High Performance Computing)

AMD SUPERCOMPUTING

OPTERON PROCESSOR AMD

AMD EPYC

AMD RADEON INSTINCT

2005
- "RED STORM" SANDIA
- "JACQUARD" NERSC
- "JAGUAR" ORNL

2006

2007
- "FRANKLIN" NERSC

2008
- "RANGER" TACC
- "ROADRUNNER" LLNL
- "KRAKEN" U OF TENNESSE/NISC

2009
- "HOPPER" NERSC

2010
- "CIELO" LLNL

2011
- US DEPT. OF ENERGY EXASCALE PROGRAM FUNDING 2012-PRESENT
- "TITAN" ORNL

2012

2019
- "CORONA" LLNL

2020
- "PERLMUTTER" NERSC

2021
- "DISCOVER" EUROHPC JU
- "CSD3" CAMBRIDGE UNIIVERSITY
- "COSMA8" DURHAM UNIVERSITY
- "LUMI" EUROHPCJU
- "FRONTIER" ORNL

2022
- UK Met Office

2023
- "EL CAPITAN" ORNL

AMD

WHO IS DEPLOYING EPYC?

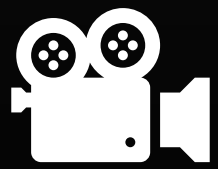
BROAD HPC ADOPTION

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE

HLRS
High-Performance Computing Center | Stuttgart

METEO FRANCE



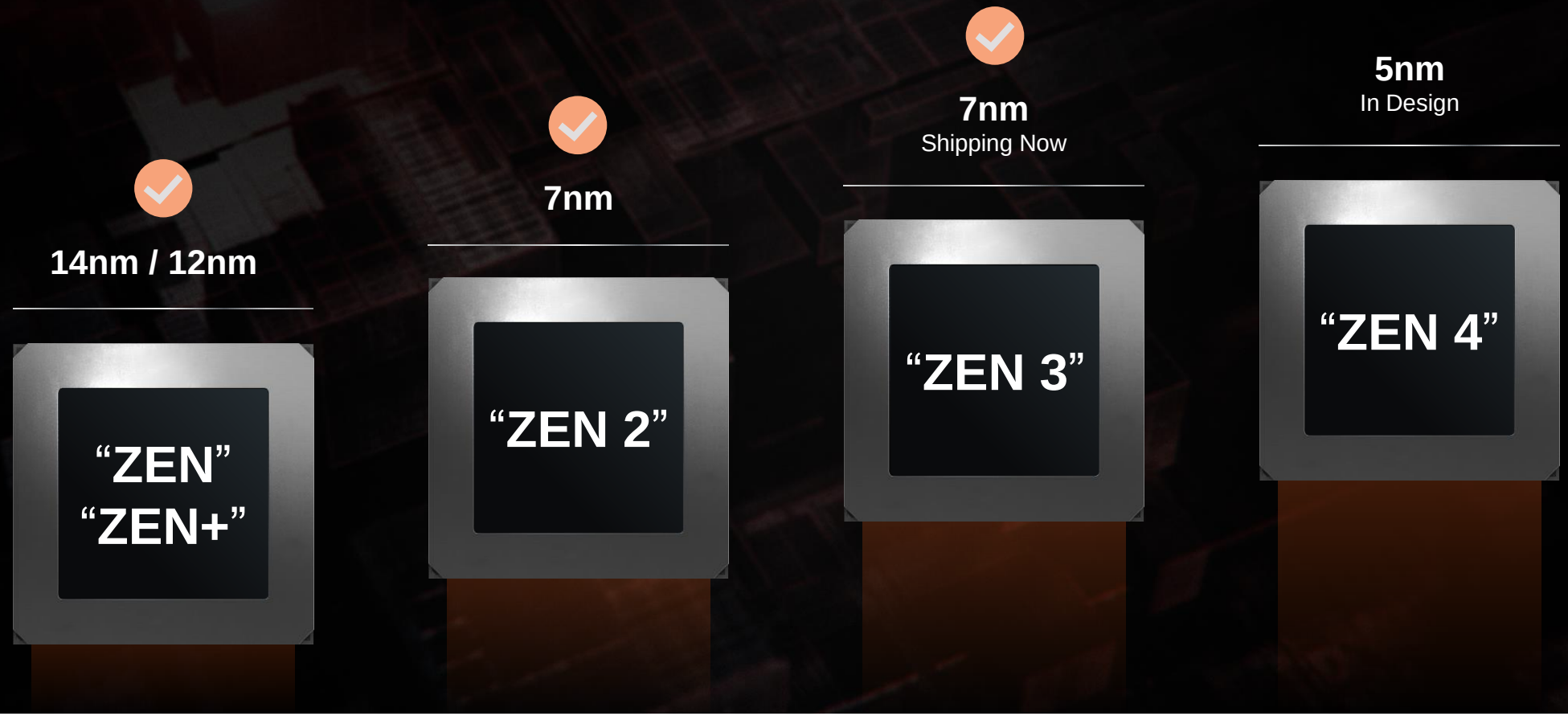
RESEARCH/ACADEMIA

COMMERCIAL

AMD RELENTLESS TECHNOLOGY INVESTMENT

CPU & GPU Leadership Roadmaps	Leading Edge Process & Packaging Technology	Next Generation Interconnect	Accelerated Computing

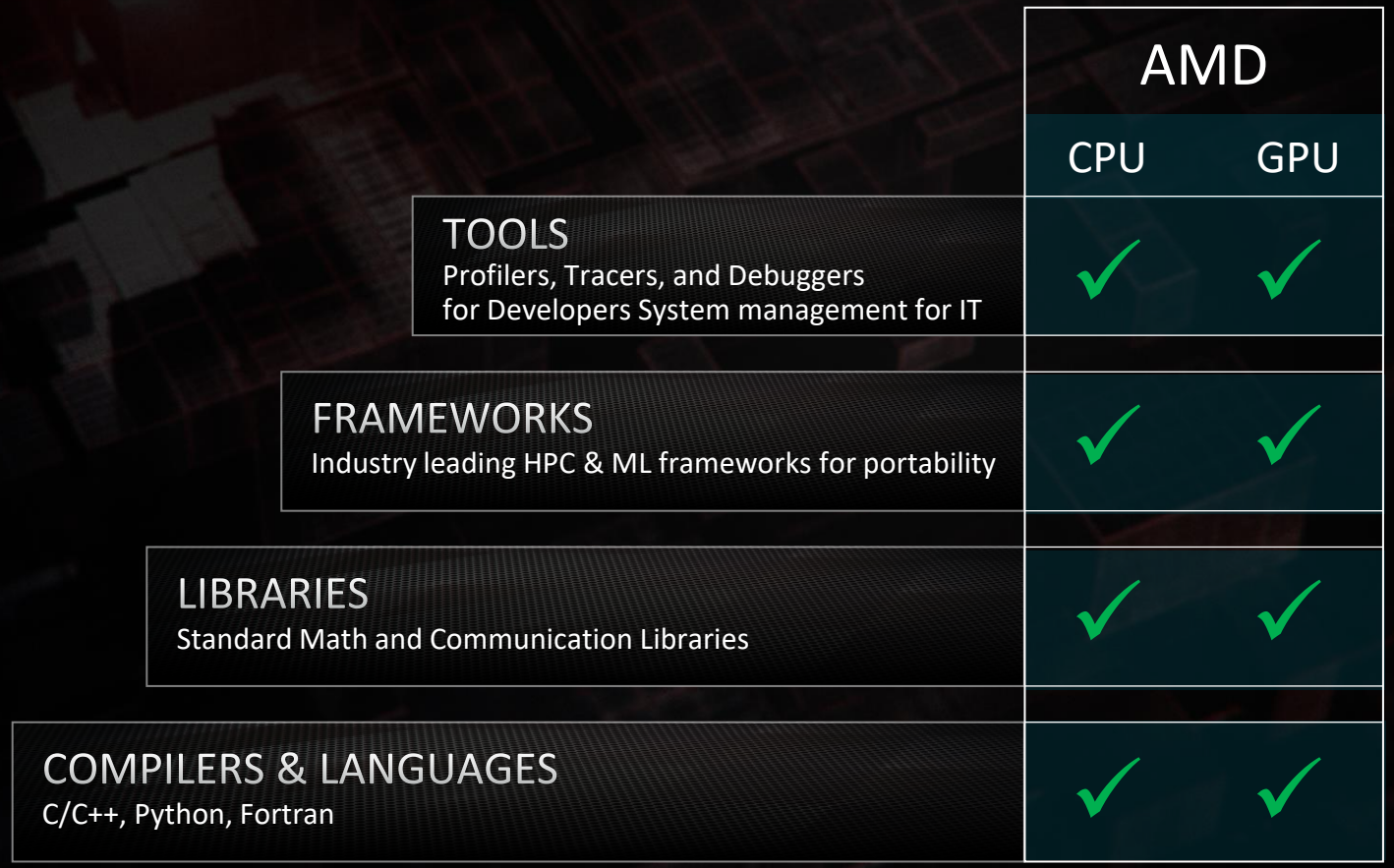
CPU ARCHITECTURE ROADMAP



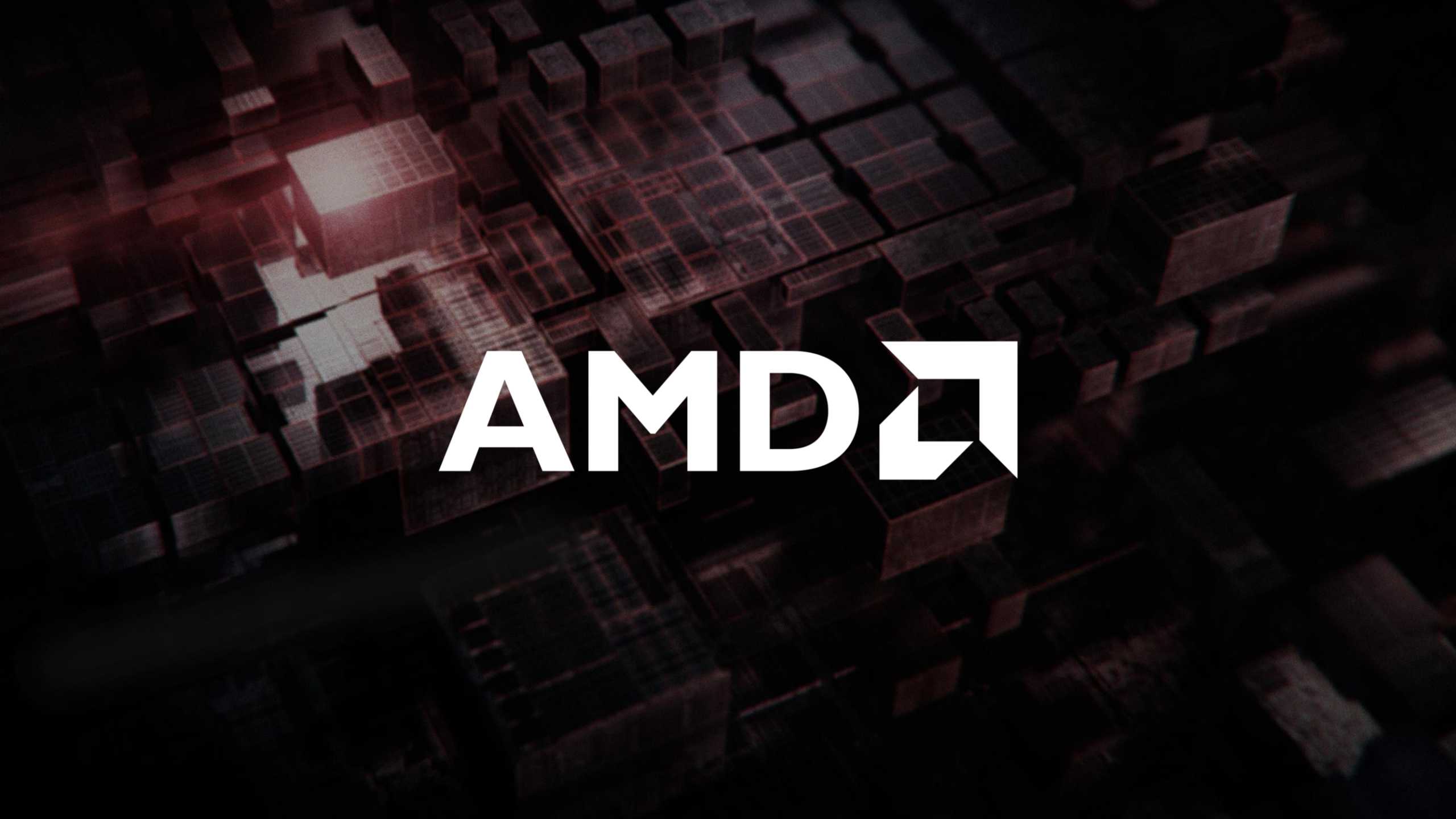
ROADMAPS SUBJECT TO CHANGE

AMD

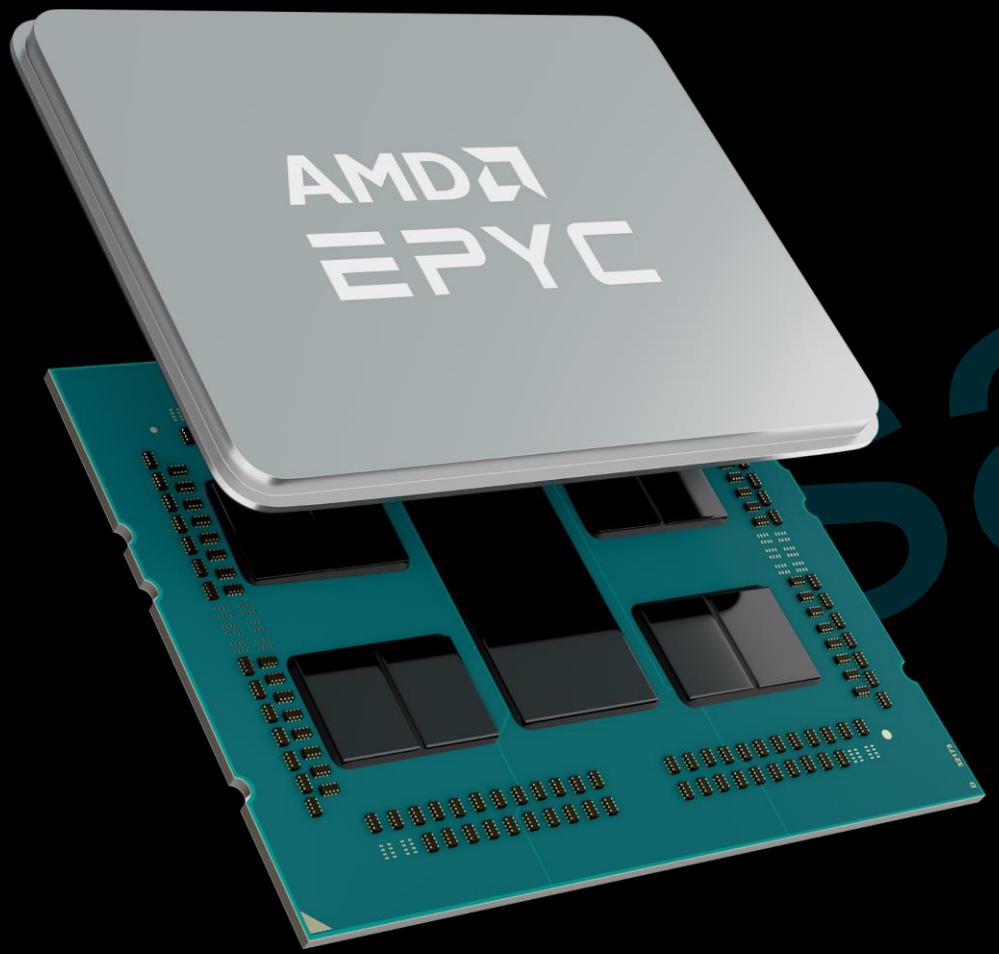
OPEN SOURCE COMMITMENT TO ENTIRE DEVELOPER STACK



AMD


AMD

KEY FEATURES FOR HPC

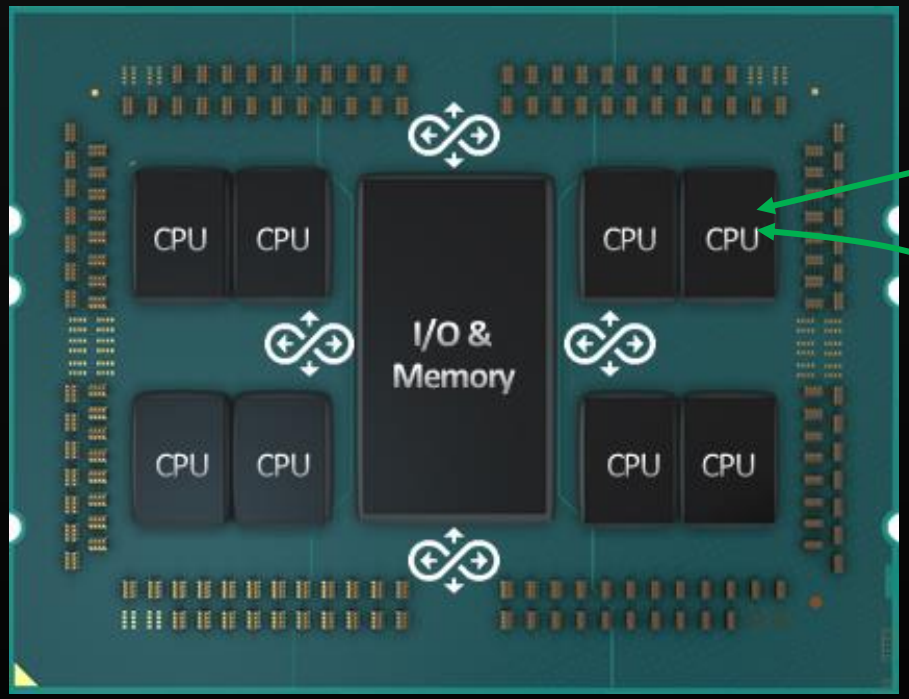


Up to 64 cores per socket	**Leadership Compute Throughput** Best choice for multi-threaded apps and capacity HPC
"Zen 3"	**Leadership core performance** Key for single-threaded apps or to maximize SW investment
8 Memory channels	**Consistent leadership since 2017** Outstanding performance for memory-bound codes
Up to 256MB L3 cache	**Most L3 cache in industry** Can enable super-linear scaling of CFD applications
Toolchain	**Commitment to Open Source** Prevent vendor lock-in

See MLN-001 and MLN-016

“MILAN” BUILDS ON INFINITY ARCHITECTURE

EPYC 7002 & 7003 DIE MCM (8 CCD + 1 IO)



7002 (“Rome”)

7003 (“Milan”)

AMD

AMD EPYC™ PROCESSORS – QUICK COMPARISON

SOCKET COMPATIBLE COMMON PLATFORM

CATEGORY	EPYC 7001	EPYC 7002	EPYC 7003
Socket	SP3	SP3	SP3 (Not Compatible With Naples MB)
Core/Process	Zen / 14nm	Zen2 / 7nm	Zen3 / 7nm
Max Core Count/Threads	32/64	64/128	64/128
L3 Cache Size	64MB	256MB	256MB
CCX Arch	4 Cores + 8MB	4 Cores + 16MB	8 Cores + 32MB
Memory	8 Ch DDR4-2666	8 Ch DDR4-3200, NVDIMM-N	8 Ch DDR4-3200, NVDIMM-N
PCIe Tech & Lane Count	PCIe Gen3, 128L/Socket	PCIe Gen4, 128L/Socket	PCIe Gen4, 128L/Socket
Security	SME, SEV	SME, SEV	SME, SEV, SNP
Chipset	NA	NA	NA
Power	120W - 180	120W - 280W	120W - 280W

AMD

AMD EPYC™ 7003 CPU ARCHITECTURE

7763 CPU - 64 CORE HIGH PERFORMANCE

AMD

AMD EPYC™ 7003 CPU ARCHITECTURE

75F3 CPU - 32 CORE HIGH PERFORMANCE / CORE

AMD

AMD EPYC™ 7003 CPU ARCHITECTURE

72F3 CPU - 8 CORE HIGH PERFORMANCE / CORE

AMD

AMD EPYC™ 7003 SERIES PROCESSOR

MODELS

ALL-IN FEATURE SET

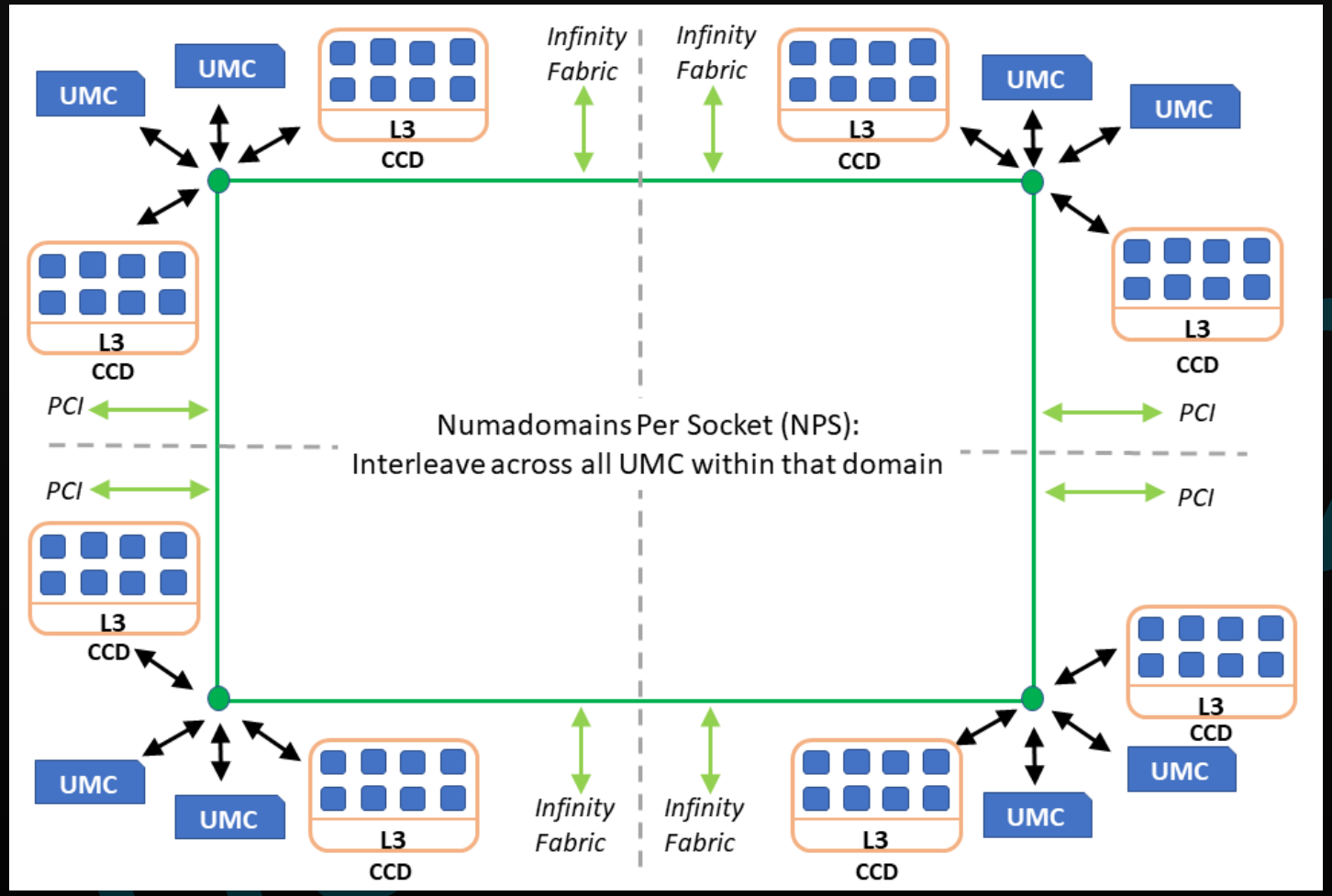
- 8 Channels of DDR4-3200
- 4TB memory capacity
- 128 lanes PCIe® 4
- SMT & Turbo boost
- 18G AMD Infinity Fabric™
- Secure Memory Encryption
- Secure Encrypted Virtualization
- Synchronized fabric and memory clock speeds

CORES	AMD EPYC	BASE/BOOST* (up to GHz)	DEFAULT TDP (W)
64 CORES	7763	**2.45/3.50**	**280W**
	7713/P	**2.00/3.675**	**225W**
56 CORES	7663	**2.00/3.50**	**240W**
48 CORES	7643	**2.30/3.60**	**225W**
32 CORES	→ 75F3	2.95/4.00	280W
	7543/P	**2.80/3.70**	**225W**
	7513	**2.60/3.65**	**200W**
28 CORES	7453	**2.75/3.45**	**225W**
24 CORES	→ 74F3	3.20/4.00	240W
	7443/P	**2.85/4.00**	**200W**
	7413	**2.65/3.60**	**180W**
16 CORES	→ 73F3	3.50/4.00	240W
	7343	**3.20/3.90**	**190W**
	7313/P	**3.00/3.70**	**155W**
8 CORES	→ 72F3	3.70/4.10	180W

"F" PERFORMANCE PER CORE OPTIMIZED *Max boost for AMD EPYC processors is the maximum frequency achievable by any single core on the processor under normal operating conditions for server systems.

AMD

MILAN ARCHITECTURE



UMC = Unified Memory Controller; CCD = Compute Core Die

COMMENTS:

- More CCDs provide increased memory BW
- PCIe and Infinity fabric assigned per NUMA. Check CTX-6 affinity
- NPS4 shown; NPS2 and NPS1 also possible
- NPS4 provides highest memory bandwidth compared to NPS2 or NPS1
- Data Fabric (DF) Speed has been increased from 1467MHz (Rome) to 1600MHz (Milan)
- 1600MHz is 'coupled' to 3200MHz, i.e. it is exactly 2 cycles in 3200. Extra memory bw increase derives from both increased DF speed and coupled mode.

AMD

```
[jason@milan003 ~]$ hwloc-ls | less     // Shows a snip from hwloc-ls //
Group0 L#2
      NUMANode L#2 (P#2 63GB)
      L3 L#4 (32MB)
        L2 L#32 (512KB) + L1d L#32 (32KB) + L1i L#32 (32KB) + Core L#32 + PU L#32 (P#32)
        L2 L#33 (512KB) + L1d L#33 (32KB) + L1i L#33 (32KB) + Core L#33 + PU L#33 (P#33)
        L2 L#34 (512KB) + L1d L#34 (32KB) + L1i L#34 (32KB) + Core L#34 + PU L#34 (P#34)
        L2 L#35 (512KB) + L1d L#35 (32KB) + L1i L#35 (32KB) + Core L#35 + PU L#35 (P#35)
        L2 L#36 (512KB) + L1d L#36 (32KB) + L1i L#36 (32KB) + Core L#36 + PU L#36 (P#36)
        L2 L#37 (512KB) + L1d L#37 (32KB) + L1i L#37 (32KB) + Core L#37 + PU L#37 (P#37)
        L2 L#38 (512KB) + L1d L#38 (32KB) + L1i L#38 (32KB) + Core L#38 + PU L#38 (P#38)
        L2 L#39 (512KB) + L1d L#39 (32KB) + L1i L#39 (32KB) + Core L#39 + PU L#39 (P#39)
      L3 L#5 (32MB)
        L2 L#40 (512KB) + L1d L#40 (32KB) + L1i L#40 (32KB) + Core L#40 + PU L#40 (P#40)
        L2 L#41 (512KB) + L1d L#41 (32KB) + L1i L#41 (32KB) + Core L#41 + PU L#41 (P#41)
        L2 L#42 (512KB) + L1d L#42 (32KB) + L1i L#42 (32KB) + Core L#42 + PU L#42 (P#42)
        L2 L#43 (512KB) + L1d L#43 (32KB) + L1i L#43 (32KB) + Core L#43 + PU L#43 (P#43)
        L2 L#44 (512KB) + L1d L#44 (32KB) + L1i L#44 (32KB) + Core L#44 + PU L#44 (P#44)
        L2 L#45 (512KB) + L1d L#45 (32KB) + L1i L#45 (32KB) + Core L#45 + PU L#45 (P#45)
        L2 L#46 (512KB) + L1d L#46 (32KB) + L1i L#46 (32KB) + Core L#46 + PU L#46 (P#46)
        L2 L#47 (512KB) + L1d L#47 (32KB) + L1i L#47 (32KB) + Core L#47 + PU L#47 (P#47)
      HostBridge
        PCIBridge
          PCI 21:00.0 (InfiniBand)
            Net "ib0"
            OpenFabrics "mlx5_0"
```

Use hwloc-ls to check

- CPU ID affinity per L3
- Location of Mellanox card

All CPU IDs in the shown NUMA domain are logically closest to the network

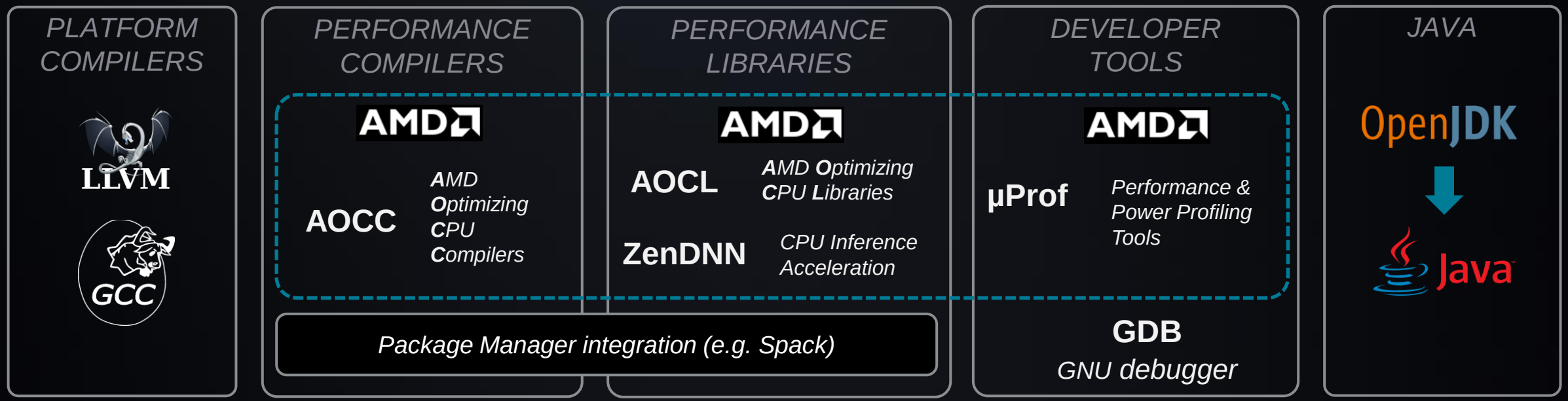
Tip: Use linux command ‘seq’ to generate lists of CPU IDs for your job, e.g. every second core:

```
[jason@milanLogin ~]$ seq -s , 0 2 127
0,2,4,6,8,10,12,14,16,18,20,22,24,26,28,30,32,34,36,38,40,42,44,46,48,50,52,54,56,58,60,62,64,66,68,70,72,74,76,78,80,82,84,86,88,90,92,94,96,98,100,102,104,106,108,110,112,114,116,118,120,122,124,126
```


AMD

TOOLCHAIN ECOSYSTEM

AMD EPYC SOFTWARE DEVELOPMENT ENVIRONMENT



Use AMD tools for best performance and code efficiency on EPYC CPUs

Compilers → focus on delivering the best out-of-the-box code generation for C, C++, Fortran, Java

Libraries → support common kernels for core math, solvers and FFT

Profiling tools → enable developers to access the full capabilities of EPYC CPUs

All tools available at developer.amd.com

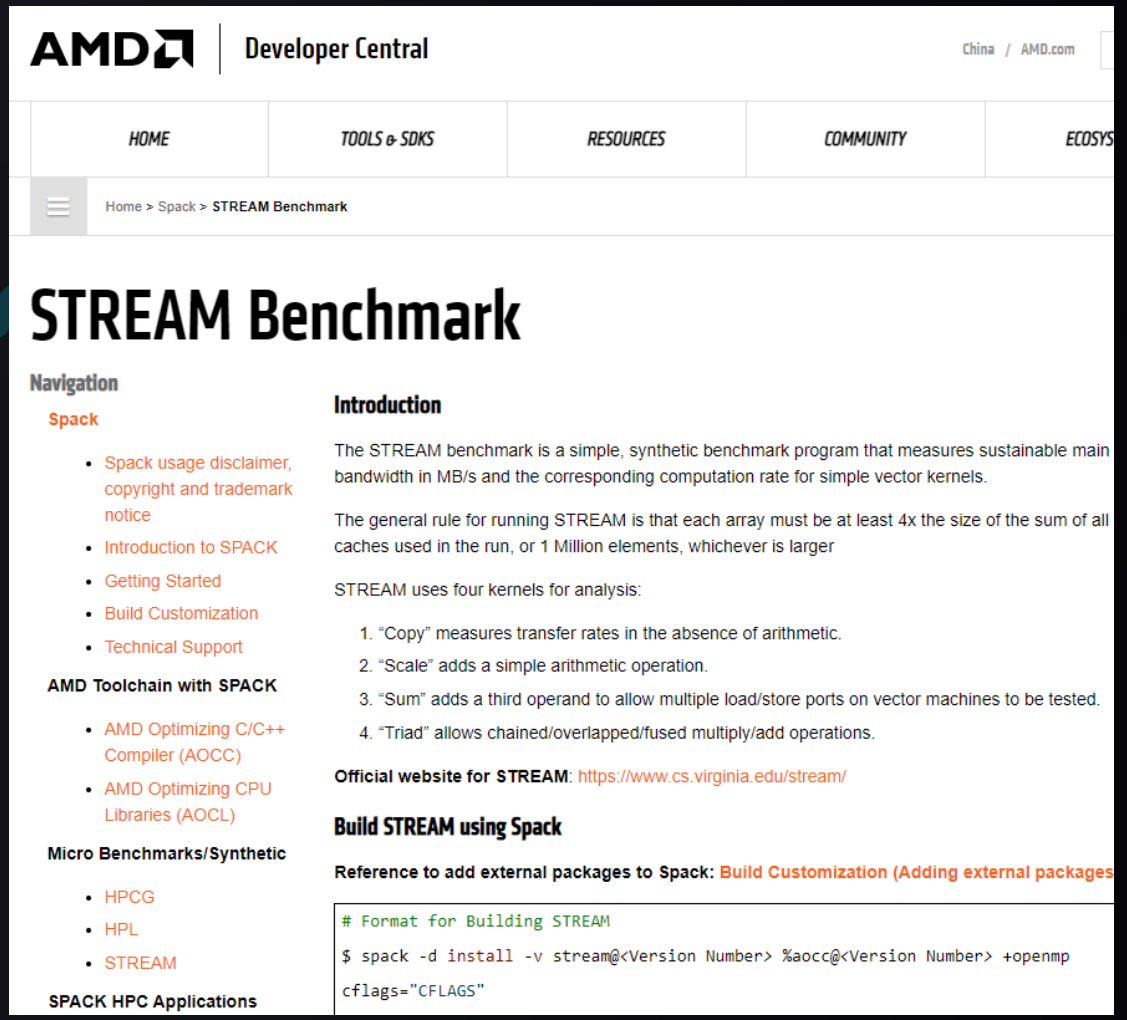
SPACK: AOCC-OPTIMISED INSTALLATION

Makes it easier for customers and partners to install and test codes and avoid extensive web searches

Install:

- AOCC +AOCL version 3.0 (aligns to “Zen3”)
- Popular synthetics, build and run steps
- Popular HPC Codes, build and run

Go to https://developer.amd.com/spack/ for application recipes (more being added, e.g. CP2K, Quantum Espresso)



AMD

SPACK: AOCC-OPTIMISED INSTALLATION

AOCL:

```
[jason@MilanLogin ~]$ spack list amd
==> 8 packages.
amdblis  amdfftw  amdlibflame  amdlibm  amdscalapack  bamdst  llvm-amdgpu  namd
```

AOCC Compiler:

```
[jason@MilanLogin ~]$ spack list aocc
==> 1 packages.
aocc
```

Example: Build Optimized Stream Binary

```
spack install stream %aocc@2.3.0 +openmp cflags="-O3 -mcmodel=large -DSTREAM_TYPE=double    \
  -mavx2 -DSTREAM_ARRAY_SIZE=2500000000 -DNTIMES=10 -ffp-contract=fast -fnt-store"

spack load stream
```


AMD

AMD EPYC™ CPUS: APPLICATIONS

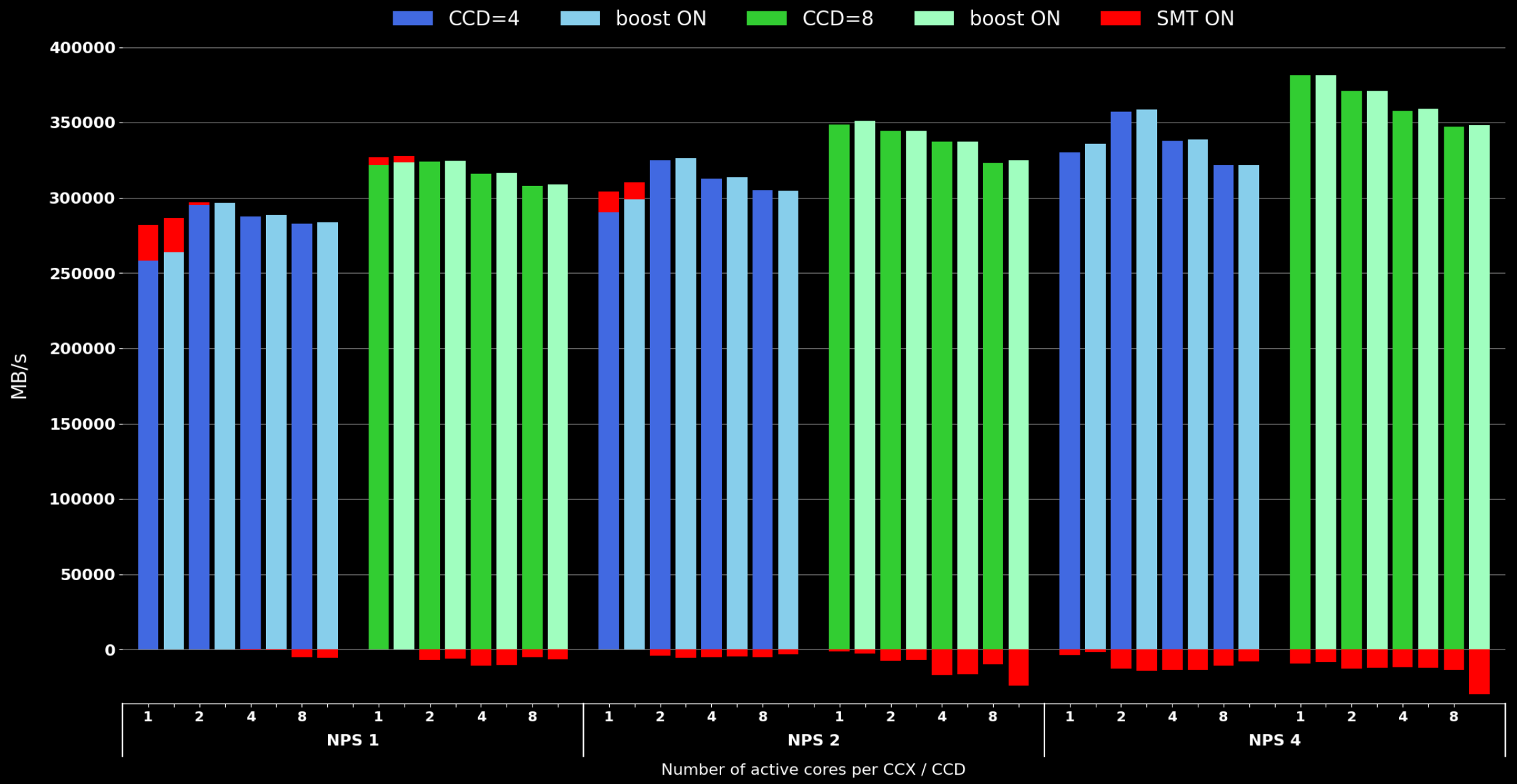
APPLICATION DOMAINS

The following domains are where AMD typically performs well in HPC.

VERY STRONG coverage across all real workloads (Life Sciences investigation starting 21Q2)

Sector	Problem Set	Example Code Set (not exhaustive)
Automotive	➢ Computational Fluid Dynamics (CFD) ➢ Structural Codes ('Finite Element Analysis' - FEA)	➢ StarCCM+, OpenFOAM, … ➢ LSDYNA, NASTRAN, ABAQUS, …
Aerospace	Same problem set as Automotive	Same code set as Automotive
Oil and Gas	➢ Reservoir ➢ Seismic	Often rely on ISV codes. Some have their own inhouse solvers
Weather	Weather Forecast and Climate Modelling	WRF, UFS, GFS, Unified Model, …
HPC Centers, including Academia	Wide range of codes: Molecular Dynamics, Quantum Chemistry, weather, CFD, inhouse codes	GROMACS, NAMD, CP2K, LAMMPS, VASP, WRF, OpenFOAM, inhouse, ….
Government Labs	Security conscious (SME?); all problem sets above may apply	Any of the above + inhouse codes

Stream Single Node Characterization

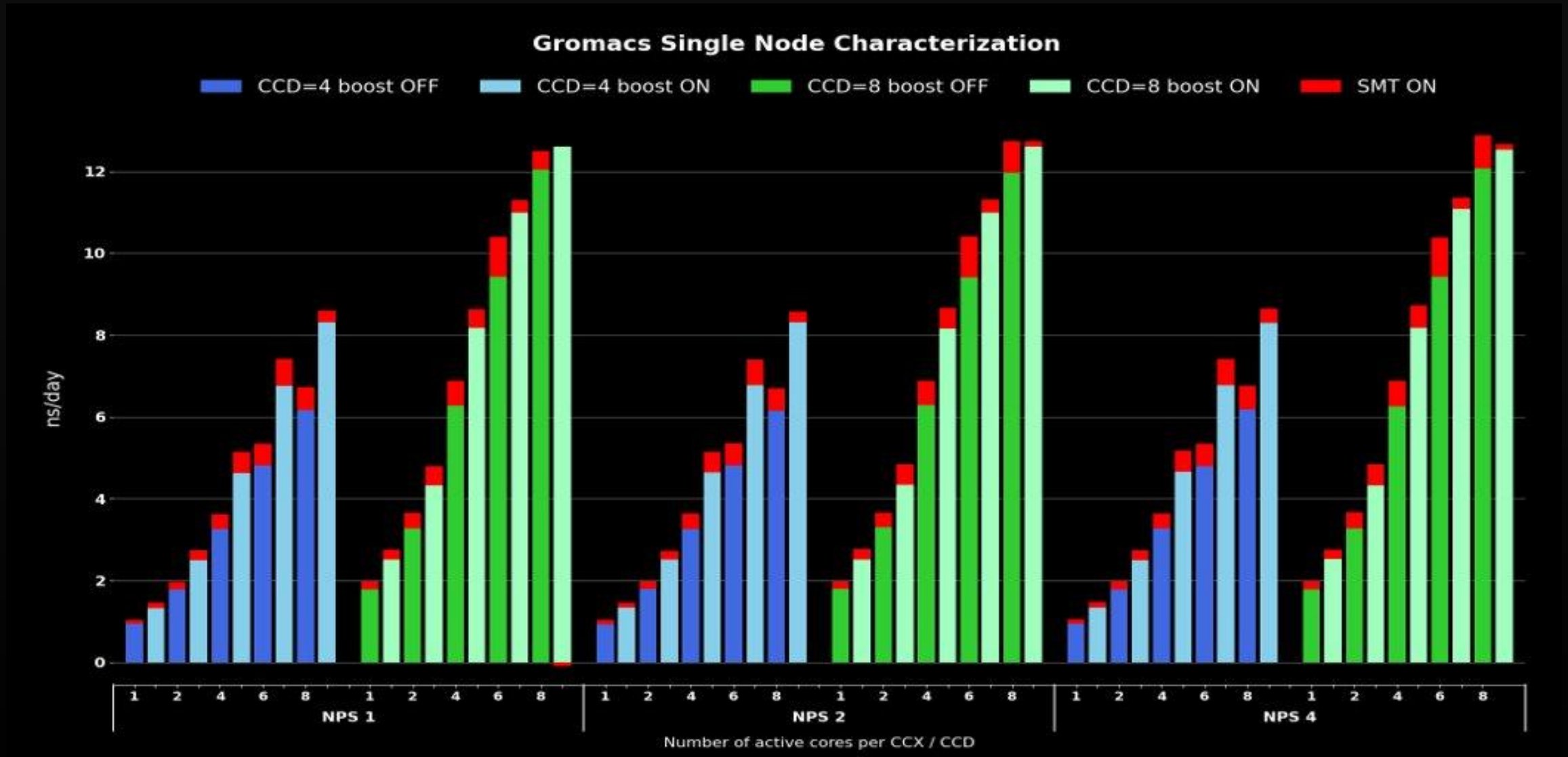


Higher is Better

CHAR-SUT-01, CHAR-APP-01

AMD

GROMACS – water_1536



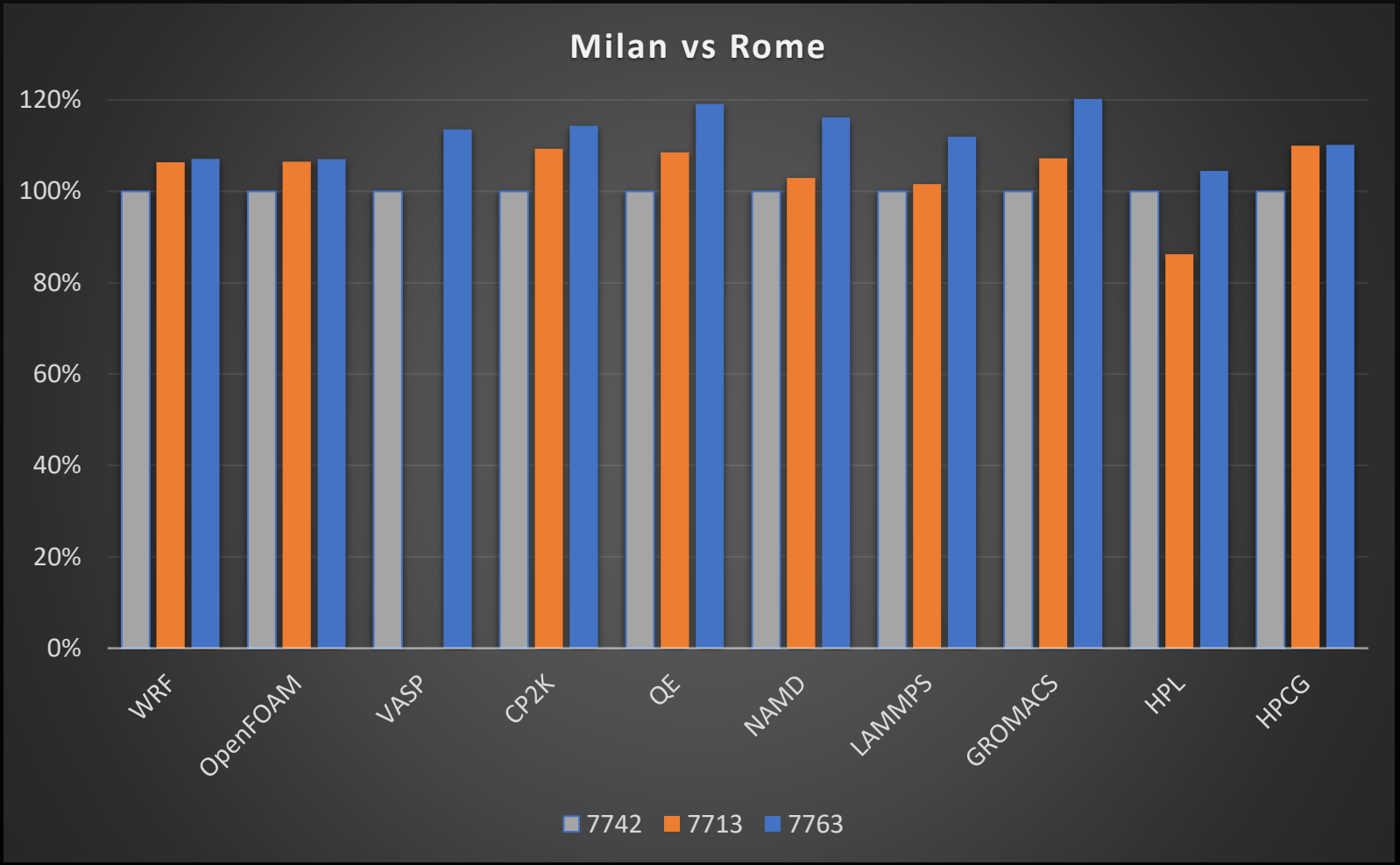
Higher is Better

CHAR-SUT-01, CHAR-APP-01

AMD

2ND GEN EPYC™ VS 3RD GEN EPYC™: COMPARING 64 CORES

Compare 7742 v 7713 v 7763



	7713	7763
WRF	6%	7%
openFOAM	7%	7%
VASP	-	14%
CP2K	9%	14%
QE	9%	19%
NAMD	3%	16%
LAMMPS	2%	12%
GROMACS	7%	22%
HPL	-14%	4%
HPCG	10%	10%

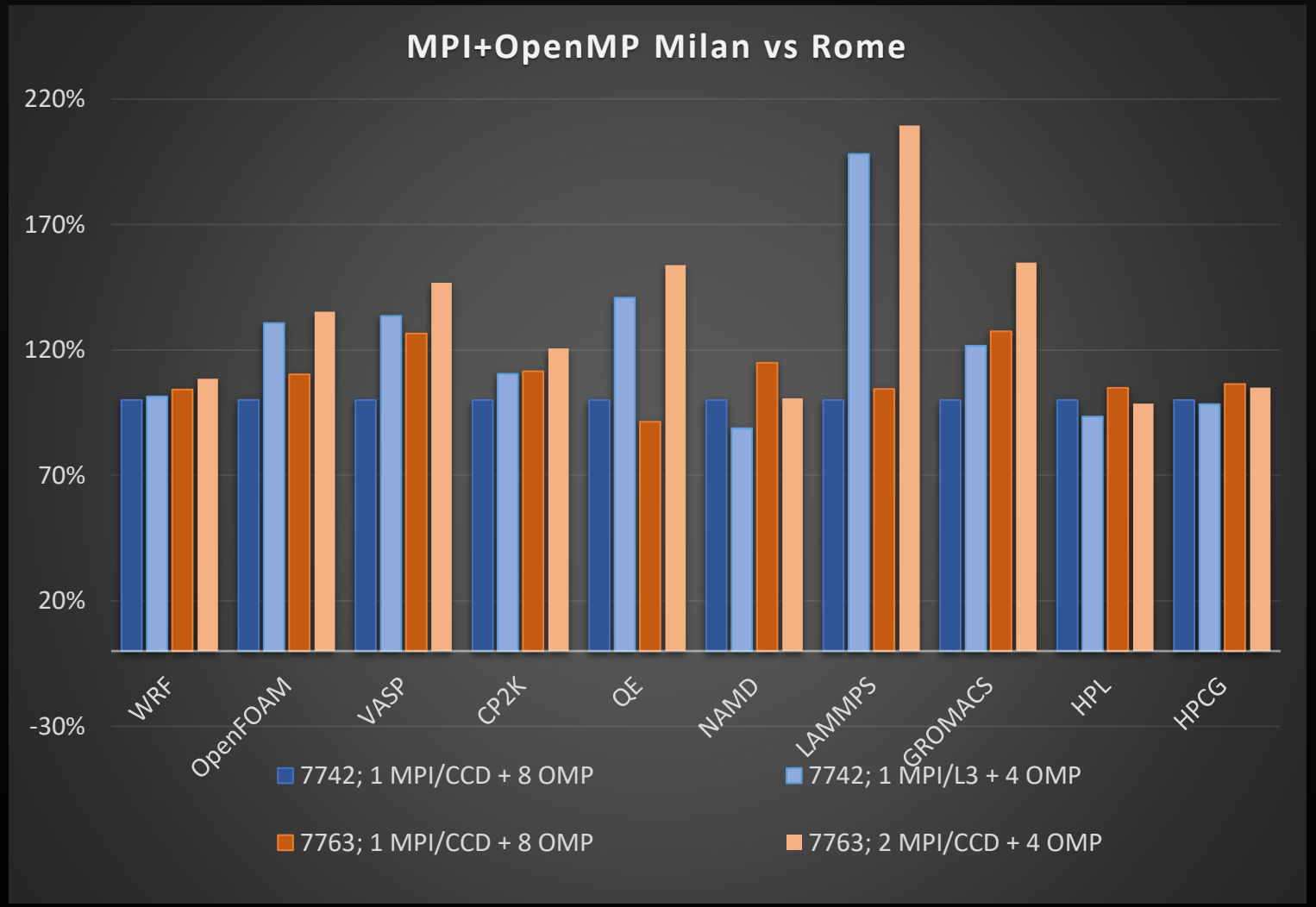
Relative to 7742

Higher is Better, Rome 7742 = 100%

CHAR-SUT-01, CHAR-APP-01

AMD

BENEFITS OF 3RD GEN EPYC™ OVER 2ND GEN EPYC™



Compare relative performance between 2nd Gen EPYC 7742 and 3rd Gen EPYC 7763 with all cores active per CDD but with differing MPI + OMP layout[Δ]:

- 7742: 1 MPI /CCD x 8 OMP *
- 7742: 1 MPI / L3 x 4 OMP
- 7763: 1 MPI / CCD x 8 OMP
- 7763: 2 MPI / CCD x 4 OMP

→ Milan improves on Rome

**Normalise to '1' and compare*

[Δ] OpenFOAM and VASP are MPI only

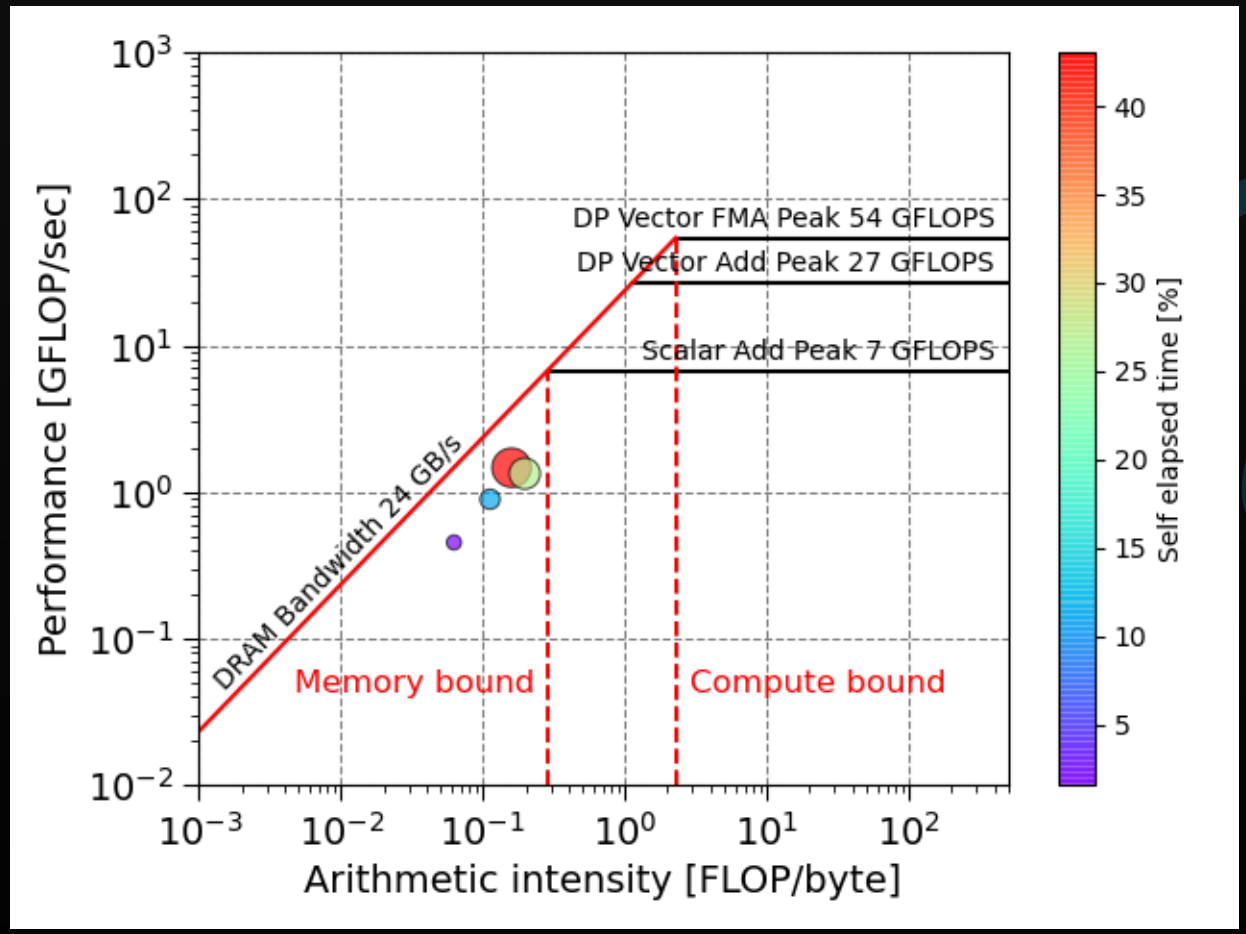
Higher is Better

CHAR-SUT-01, CHAR-APP-01

AMD

MEMORY BANDWIDTH v CLOCK FREQ

LAMMPS, EMA



2 competing system-level choke-points:

- Bandwidth to main memory
- 'Compute Bound ('frequency')

These are mutually exclusive to each other

Perform roofline analysis to confirm where hot-routine lands (red circle)

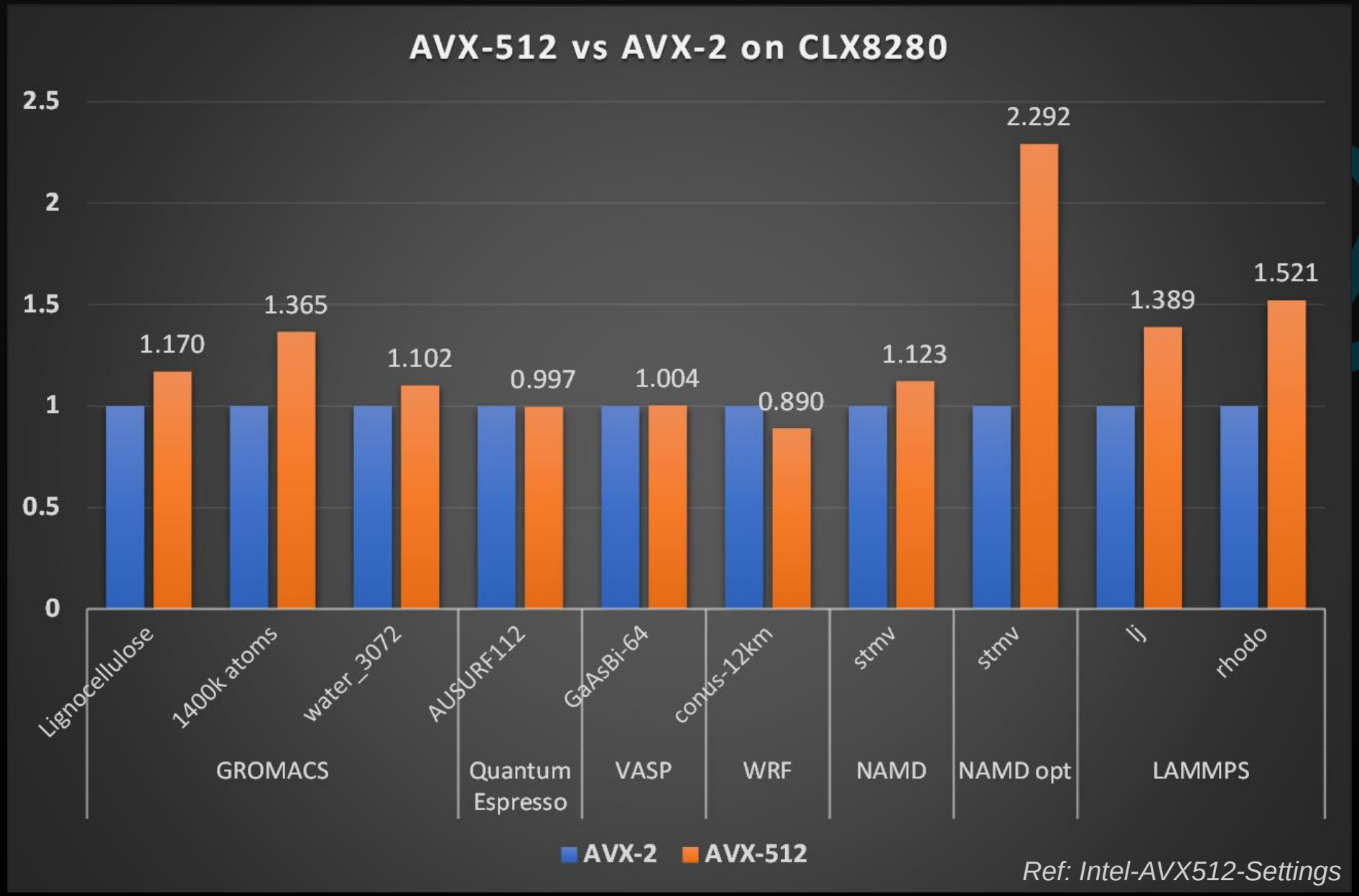
We have performed this analysis on a number of popular HPC codes across CFD, Weather, Quantum Chemistry, Molecular Dynamics: Codes are memory bound or borderline

→ HPL (compute bound) is *NOT* a good proxy for scoping job-throughput on realistic workloads.

→ Use memory-bound synthetics: HPCG or STREAM

Footnote: CHAR-SUT-02, single core

ADVANCED VECTOR EXTENSIONS



Ref: Intel-AVX512-Settings

Some analysis undertaken within Centre of Excellence. Used Intel system:

- Compiled with AVX2
- Compiled with AVX512

HPL is ‘compute bound’

Theoretically;
AVX512 = 2x AVX2

Some improvement over AVX2 …

… but extra core count on EPYC™ CPUs will make up difference in performance

HPC TOOLS AND RESOURCES

Developer Hub:

http://developer.amd.com

- Download compilers, math libraries, uProf profiling tool
- User guides

AMD Tech Docs & White Papers Site:
https://www.amd.com/en/processors/server-tech-docs/search?f%5B0%5D=product_type%3Aepyc_7003&f%5B1%5D=server_document_category%3A14241

- Solution Briefs
- Performance Briefs

AMD HPC Tuning Guide for 7003 ‘Milan’ Processors
https://www.amd.com/system/files/documents/high-performance-computing-tuning-guide-amd-epyc7003-series-processors.pdf

TECHNOLOGY FOR THE FUTURE

Roadmap for Relentless Compute Efficiency Gains	Aggressive Process & Packaging Deployment	Faster Time to Market Through Innovative Execution	Exascale High-Performance Driving Next Wave of Innovation

AMD


AMD

CHAR-SUT-01:

System	Gigabyte MZ62-HD0-00
BIOS version	M04 (AGESA 1.0.0.2)
Operating System	CentOS 8.3
Kernel	4.18.0-240.22.1.el8_3.x86_64
OFED	5.2.2.2
Memory	16 channels x DDR4-3200 32GB R2

For 2nd Gen EPYC™ v 3rd Gen EPYC™ comparisons, the above System config was used with either a 2nd Gen EPYC™ 7742 CPU or 3rd Gen EPYC™ 7713/7763 CPU as appropriate

BIOS was set per the HPC Tuning Guide, and altered accordingly for the tests herein

---

CHAR-SUT-02 YYY UPDATE):

System	'Daytona'
CPU	7742, Rome
BIOS version	5.14, 08/15/2019
Operating System	CentOS 7.6
Kernel	3.10.0-957.10.1.el7.x86_64
OFED	MLNX_OFED_LINUX-4.7-3.2.9.0
Memory	16 channels x DDR4-3200 32GB R2

SMT=OFF,, Boost=ON, C2 disabled, Determinism Slider=Power, CPU governor=Performance

CHAR-APP-01

LAMMPS, 29-OCT-20 lj
AOCC+AOCL: -g -O3 -march=znver2 -ffast-math -ffp-contract=fast -DNDEBG -pthread -fopenmp -std=c++11

VASP, 6.1.2 GaAsBi-64 *(VASP guide refers to –02 and not –O3 as default optimisation level)*
Intel: -O2 -assume byterecl -w -march=core-avx2 -fma -mavx2 -mfma -free -names lowercase -mkl=sequential -qopenmp -lmkl_blacs_openmpi_lp64 -lmkl_scalapack_lp64

CP2K, 7.1 H20-256
AOCC+AOCL compile Line: -O3 -march=znver3 -ffast-math -fopenmp -ffree-form -Hx,47,0x10000008 -Mbackslash

Quantum Espresso, 6.5 AUSURF112
GCC+AOCL compile Line: -O3 -march=znver2 -ffast-math -fopenmp

WRF, 3.9.1.1, CONUS 2.5KM
AOCC: -march=znver3 -ffp-contract=fast -mavx2 -mfma -mprefer-vector-width=256 -Ofast -ffast-math -fopenmp -m64 -w -finline-aggressive -fPIC

OpenFOAM, 2012, Motorbike 42m cells
GCC: -O3 -march=znver2 -mfma -ffp-contract=fast -funroll-loops -fno-signed-zeros -fno-trapping-math -fno-math-errno -ffinite-math-only -fno-rounding-math -fno-signaling-nans -fcx-limited-range -fexcess-precision=fast

GROMACS, 2020.4 water_1536
AOCC+AOCL: -DGMX_DOUBLE=0 -O3 -Ofast -g -m64 -march=znver3 -fPIC -ffast-math -ffp-contract=fast -Wno-deprecated-copy -Wno-error=deprecated-copy
--gcc-toolchain=/usr -O3 -DNDEBUG -mavx2 -mfma -funroll-all-loops -pthread -fopenmp=libomp

NAMD, 2.14
AOCC+AOCL: -optimize -production -DCMK_OPTIMIZE -O3 -Ofast -g -m64 -march=znver3 -fPIC -ffp-contract=fast -fopenmp -Wno-deprecated-register
-Wno-non-c-typedef-for-linkage -mllvm -inline-threshold=3000 -finline-hint-functions -mllvm -inlinehint-threshold=10000 -finline-functions

Intel-AVX512-Settings

CPU	Intel(R) Xeon(R) Platinum 8280 CPU @ 2.70GHz, 28 cores, 2 sockets
Motherboard	Supermicro X11DPT-B 1.02
Memory	DDR4 2933MT/s, 376GiB
BIOS	American Megatrends Inc. v3.1 04/30/2019
Kernel	4.18.0-193.el8.x86_64
CPU frequency	Performance governor, intel_pstate driver, boost active
Hyperthreading	On
Compiler	ICC/IFORT 19.1.2.254 20200623
Math Library	Intel MKL 2020.2.254
AVX2 flags	-O3 -xCORE-AVX2 MKL_ENABLE_INSTRUCTIONS=AVX2
AVX512 flags	-O3 -xCORE-AVX512 -qopt-zmm-usage=high
Codes versions	GROMACS v2021.1 Quantum Espresso v6.7 VASP v6.2.0 WRF v4.2.1 NAMD 2.14, NAMD v2.15alpha1 LAMMPS stable_29Oct2020

END NOTES

MLN-073K: Based on SPECrate®2017_int_base on 02/20/2021, a server powered by two 64c AMD EPYC 7763 CPUs has a score of 819 in a compliant result run on an ThinkSystem SR645; with Memory: 2 TB (32 x 64 GB 2Rx4 PC4-3200AA-R), OS: SUSE Linux Enterprise Server 12 SP5 (x86_64) Kernel 4.12.14-120-default; Compiler: C/C++/Fortran: Version 3.0.0 of AOCC. Versus the current highest score Intel Cascade Lake Refresh server with a score of 397 using 2P Intel Gold 6258R, https://spec.org/cpu2017/results/res2020q3/cpu2017-20200915-23981.pdf. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

MLN-074K: Based on SPECrate®2017_fp_base on 02/20/2021, a server powered by two 64c AMD EPYC 7763 CPUs has a score of 636 a compliant result run on an ThinkSystem SR665; with Memory: 512 GB (16 x 32 GB 2Rx4 PC4-3200AA-R); OS: Red Hat Enterprise Linux release 8.3 (Ootpa); Compiler: C/C++/Fortran: Version 3.0.0 of AOCC. Versus the current highest score Intel Cascade Lake Refresh server with a score of 309 with a 2P Intel Gold 6258R based server, https://spec.org/cpu2017/results/res2020q3/cpu2017-20200915-23979.pdf. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

MLN-044A: SPECjbb®2015-MultiJVM critical-jOPS comparison based on highest system results published as of 03/11/2021. Configurations: 2x AMD EPYC 7763 (301297 SPECjbb2015-MultiJVM critical-jOPS, 359067 SPECjbb2015-MultiJVM max-jOPS, https://spec.org/jbb2015/results/res2021q1/jbb2015-20210224-00610.html) versus 2x Intel Xeon Platinum 8280 (138942 SPECjbb2015-MultiJVM critical-jOPS, 169,598 SPECjbb2015-MultiJVM max-jOPS, https://spec.org/jbb2015/results/res2019q2/jbb2015-20190314-00428.html) for 117% higher [~2.2x the] performance. SPEC®, and the benchmark SPECjbb® are registered trademarks of the Standard Performance Evaluation Corporation. Learn more at spec.org.

MLN-001: AMD EPYC™ 7003 Series processors require a BIOS update from your server or motherboard manufacturer if used with a motherboard designed for the AMD EPYC™ 7002 Series processors. A motherboard designed at minimum for EPYC 7002 processors is required for EPYC 7003 Series processors.

MLN-016: Results as of 01/28/2021 using SPECrate®2017_int_base. The AMD EPYC™ 7763 a measured estimated score of 798 is higher than the current highest 2P server with an AMD EPYC™ 7H12 and a score of 717, https://spec.org/cpu2017/results/res2020q2/cpu2017-20200525-22554.pdf. OEM published score(s) for 3rd Gen AMD EPYC™ may vary. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

MLN-071K: Based on SPECrate®2017_int_base on 02/20/2021, a server powered by two 64c AMD EPYC 7763 CPUs has a score of 839 which is higher than any currently posted SPEC 2P server score. Per socket score would be 839/2=419.5 which is higher than any 1P server score. This is a compliant result run on an ASUS RS720A-E11(KMPP-D32); with Memory: 1 TB (16 x 64 GB 2Rx4 PC4-3200AA-R); OS: SUSE Linux Enterprise Server 15 SP2 (x86_64) Kernel 5.3.18-22-default; Compiler: C/C++/Fortran: Version 3.0.0 of AOCC. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

EPYC-22: For a complete list of world records see http://amd.com/worldrecords.

AMD

END NOTES CONTINUED

MLN-057K: Based on SPECrate®2017_int_base on 02/20/2021, a server powered by two 8c AMD EPYC™ 72F3 CPU has a measured estimated score of 176 with a per core score of 11.00 which is a higher per core performance score than any currently posted in any SPEC.org publication. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

EPYC-22: For a complete list of world records see http://amd.com/worldrecords.

ROM-693: 2nd Gen EPYC™ 7H12 powered server with 2 total processors (1-node, 2-socket) scoring 248,942 SPECjbb®2015-MultiJVM critical-jOPS (315,663 max-jOPS, http://spec.org/jbb2015/results/res2020q2/jbb2015-20200423-00550.html) has up to 1.79x the SPECjbb® 2015-MultiJVM Critical performance of the highest competitive score of 138942 SPECjbb®2015-MultiJVM critical-jOPS (169598 max-jOPS, http://www.spec.org/jbb2015/results/res2019q2/jbb2015-20190314-00428.html) on a Xeon® Platinum 8280 powered server with 2 total processors (1-nodes, 2-socket) as of 6/9/20. SPEC® and SPECjbb® are trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information.

CDNA-02: The AMD Instinct™ MI100 accelerator has 120 compute units (CUs) and 7,680 stream cores at 300W. The Radeon Instinct™ MI50 GPU has 60 CUs and 3,840 stream cores at 300W. CDNA-02

CDNA-04 : Calculations by AMD Performance Labs as of Oct 5th, 2020 for the AMD Instinct™ MI100 accelerator designed with AMD CDNA 7nm FinFET process technology at 1,200 MHz peak memory clock resulted in 1.2288 TFLOPS peak theoretical memory bandwidth performance. The results calculated for Radeon Instinct™ MI50 GPU designed with “Vega” 7nm FinFET process technology with 1,000 MHz peak memory clock resulted in 1.024 TFLOPS peak theoretical memory bandwidth performance. CNDA-04

MI100-04: Calculations performed by AMD Performance Labs as of Sep 18, 2020 for the AMD Instinct™ MI100 accelerator at 1,502 MHz peak boost engine clock resulted in 184.57 TFLOPS peak theoretical half precision (FP16) and 46.14 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. The results calculated for Radeon Instinct™ MI50 GPU at 1,725 MHz peak engine clock resulted in 26.5 TFLOPS peak theoretical half precision (FP16) and 13.25 TFLOPS peak theoretical single precision (FP32 Matrix) floating-point performance. Server manufacturers may vary configuration offerings yielding different results. MI100-4

AMD

END NOTES CONTINUED

M100-06: As of SEP 18th, 2020. AMD Instinct™ MI100 built on AMD CDNA technology accelerators support PCIe® Gen4 providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card. Peak theoretical transport rate performance is calculated by Baud Rate * width in bytes *# directions = GB/s per card. PCIe Gen4: 16 * 2 * 2 = 64 GB/s. AMD Instinct™ MI100 "CDNA" technology-based accelerators include three Infinity Fabric™ link providing up to 276 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) transport rate bandwidth performance per GPU card. Combined with PCIe Gen4 compatibility providing an aggregate GPU card I/O peak bandwidth of up to 340 GB/s. Infinity Fabric link technology peak theoretical transport rate performance is calculated by Baud Rate * width in bytes *# directions *# links = GB/s per card. Infinity Fabric Link: 23 * 2 * 2 = 92 GB/s. MI100s have three links: 92 GB/s * 3 links per GPU = 276 GB/s. Four GPU hives provide up to 552 GB/s peak theoretical P2P performance: 276GB/s * 4 GPUs = 1,104 GB/s, divided by 2 = 552 GB/s. Dual 4 GPU hives in a server provide up to 1.1 TB/s total peak theoretical direct P2P performance per server: 552 GB/s per 4 GPU hive * 2 = 1,104 GB/s. AMD Infinity Fabric link technology not enabled: Four GPU hives provide up to 256 GB/s peak theoretical P2P performance with PCIe® 4.0: 64GB/s per GPU * 4 GPUs = 256 GB/s. Radeon Instinct™ MI50 "Vega 7nm" technology-based accelerators support PCIe® Gen 4.0* providing up to 64 GB/s peak theoretical transport data bandwidth from CPU to GPU per card. Radeon Instinct™ MI50 "Vega 7nm" technology-based accelerators include dual Infinity Fabric™ Links providing up to 184 GB/s peak theoretical GPU to GPU or Peer-to-Peer (P2P) transport rate bandwidth performance per GPU card. Combined with PCIe Gen 4 compatibility providing an aggregate GPU card I/O peak bandwidth of up to 248 GB/s. MI50 based four GPU hives provide up to 368 GB/s peak theoretical P2P performance: 184 GB/s * 4 GPUs = 736 GB/s, divided by 2 = 368 GB/s. Dual 4 GPU hives in a server provide up to 736 GB/s total peak theoretical direct P2P performance per server: 368 GB/s per 4 GPU hive * 2 = 736 GB/s. Performance guidelines are estimated only and may vary. Refer to server manufacture PCIe Gen4 compatibility and performance guidelines for potential peak performance of the specified server model numbers. Server manufacturers may vary configuration offerings yielding different results. https://pcisig.com/, https://www.chipestimate.com/PCI-Ex- press-Gen-4-a-Big-Pipe-for-Big-Data/Cadence/Technical-Article/2014/04/15, https://www.tomshardware.com/news/pcie-4.0-power-speed-express,32525.html AMD has not independently tested or verified external/third party results/data and bears no responsibility for any errors or omissions therein. MI100-06.

AMD

END NOTES CONTINUED

MLNTCO-001: The Bare Metal TCO (total cost of ownership) Estimator solution compares the selected AMD EPYC™ and Intel® Xeon® CPU based server solutions required to deliver a TOTAL_PERFORMANCE of 25000 unit of integer performance based on published the SPECrate®2017_int_base scores for Intel and AMD measured estimated scores for AMD EPYC 7003. This analysis is based on tool VERSION: 02/20/2021 v0.9982. This estimation reflects a 4 year time frame.

This analysis compares a 2 CPU AMD EPYC EPYC_7763 powered server with a measured estimated SPECrate®2017_int_base score of 802; compared to a 2 CPU Intel Xeon Gold_6258R based server with a SPECrate®2017_int_base score of 397, https://spec.org/cpu2017/results/res2020q3/cpu2017-20200915-23981.pdf.

Both AMD EPYC and Intel based servers use the same estimated cost for the following elements of the analysis: server chassis size of 2RU at a cost of $2500 per chassis; internal storage $380; physical servers managed per admin: 30; fully burdened cost per admin $110500; server rack size of 42; space allowance per rack of 27 sq feet; monthly cost of data center space $20 per sq foot; cost per kW for power $0.12; power drop per rack of 12kW; and a PUE (power usage effectiveness of 2).

The EPYC powered solution estimates are: 32 2P EPYC 7763 powered total servers at a hardware only acquisition cost of $19232 per server, which includes total system memory of 768GB, which is 6GB of memory / core and a total system memory cost of $3072; internal storage cost of $380. The total AMD EPYC hardware acquisition cost for this solution is $615424. Each server draws ~611kWhr per month. For the 4 years of this EPYC powered solution analysis the: total solution power cost is ~$225240 which includes the PUE factor; the total admin cost is ~$471468, and the total real estate cost is ~$77760. The total 4 year TCO estimate for the AMD solution is $1389892.

The Intel based solution estimates are: 63 2P Xeon Gold 6258R based total servers at a hardware only acquisition cost of $12316 per server, which includes total system memory of 384GB, which is 6.9GB of memory / core and a total system memory cost of $1536; internal storage cost of $380. The total Intel hardware acquisition cost for this solution is $775908. Each server draws ~476kWhr per month. For the 4 years of this Intel based solution analysis the: total solution power cost is $345460 which includes the PUE factor; the total admin cost is ~$928200, and the total real estate cost is ~$103680. The total 4 year TCO estimate for the Intel solution is $2153248.
Delivering 25000 of estimated SPECrate®2017_int_base performance, produces the following estimated results: the AMD EPYC solution requires 49% fewer servers [1-(AMD server count / Intel server count)]; 25% less space [1-(AMD rack count / Intel rack count)]; 35% less power [1-(AMD power cost / Intel power cost)]; providing a 35% lower 4 year TCO [1-(AMD TCO / Intel TCO)].

AMD processor pricing based on 1KU price as of February 2021. Intel® Xeon® Scalable processor data and pricing from https://ark.intel.com as of September 2020. All pricing is in USD. Results shown here are estimates and actual results may vary. Product and company names are for informational purposes only and may be trademarks of their respective owners. SPECrate® scores as of 02/20/2021. AMD EPYC performance numbers based on AMD internal estimates and are subject to change based on actual results. SPEC®, SPECrate® and SPEC CPU® are registered trademarks of the Standard Performance Evaluation Corporation. See www.spec.org for more information. AMD EPYC performance numbers based on AMD measured internal estimates and are subject to change based on actual results. Results generated by the AMD EPYC™ BARE METAL SERVER TCO ESTIMATION TOOL, VERSION: 02/20/2021 v0.9982.

AMD

Disclaimer and Attributions

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. The information contained herein is subject to change and may be rendered inaccurate for many reasons, including but not limited to product and roadmap changes, component and motherboard version changes, new model and/or product releases, product differences between differing manufacturers, software changes, BIOS flashes, firmware upgrades, or the like. Any computer system has risks of security vulnerabilities that cannot be completely prevented or mitigated. AMD assumes no obligation to update or otherwise correct or revise this information. However, AMD reserves the right to revise this information and to make changes from time to time to the content hereof without obligation of AMD to notify any person of such revisions or changes.

THIS INFORMATION IS PROVIDED ‘AS IS.” AMD MAKES NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE CONTENTS HEREOF AND ASSUMES NO RESPONSIBILITY FOR ANY INACCURACIES, ERRORS, OR OMISSIONS THAT MAY APPEAR IN THIS INFORMATION. AMD SPECIFICALLY DISCLAIMS ANY IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMD BE LIABLE TO ANY PERSON FOR ANY RELIANCE, DIRECT, INDIRECT, SPECIAL, OR OTHER CONSEQUENTIAL DAMAGES ARISING FROM THE USE OF ANY INFORMATION CONTAINED HEREIN, EVEN IF AMD IS EXPRESSLY ADVISED OF THE POSSIBILITY OF SUCH DAMAGES.

AMD, the AMD Arrow logo, EPYC, and combinations thereof are trademarks of Advanced Micro Devices, Inc. PCIe is a registered trademark of PCI-SIG Corporation. Other product names used in this publication are for identification purposes only and may be trademarks of their respective companies.